DOCUMENT RESUME

ED 194 604 TM 800 735

TITLE	A Guide for Local District Testing Programs. Occasional Papers in Education.
INSTITUTION	New Jersey State Dept. of Education, Trenton. Div. of Operations, Research, and Evaluation.
PUB DATE	80
NOTE	64p.
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	Elementary Secondary Education; *Guidelines; Program Evaluation; *School Districts; Scores; Test Construction; *Testing Programs; Test Interpretation; Test Results
IDENTIFIERS	*New Jersey

ABSTRACT

Local educational agencies and county office staff are provided with information and specific guidelines on the development, maintenance and evaluation of local district testing programs. This manual contains New Jersey approved criteria for local testing programs, a step-by-step presentation of methods for planning, designing and evaluating testing programs, summary information about tests, scoring procedures, types of scores, methods of reporting and interpreting results and a selective annotated reference section. A commercial test selection checklist is presented, as well as a discussion on the advantages and disadvantages associated with locally developed tests. Assistance in the development of local tests is available from the Office of State Educational Assessment or the Educational Improvement Centers. Information is provided about paper and pencil tests only. (RL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED194604

Occasional Papers in Education

NEW JERSEY STATE DEPARTMENT OF EDUCATION

DIVISION OF OPERATIONS, RESEARCH, AND EVALUATION

A GUIDE FOR LOCAL DISTRICT TESTING PROGRAMS

DEC 1980 RECEIVED ERIC/TM

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

L F. Price

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

A GUIDE FOR LOCAL DISTRICT TESTING PROGRAMS

Fred G. Burke
Commissioner of Education

Gustav H. Ruh
Deputy Commissioner of Education

PREPARED UNDER THE DIRECTION OF:

OFFICE OF STATE EDUCATIONAL ASSESSMENT
DIVISION OF OPERATIONS, RESEARCH AND EVALUATION

Arthur Winkler
Assistant Commissioner
Division of Operations, Research and Evaluation

WRITTEN AND EDITED BY:

Office of State Educational Assessment

New Jersey State Department of Education
225 West State Street
P.O. Box 2019
Trenton, New Jersey 08625

1980

P.T.M. NO. 100.60

0 3

DEC 4 1980

STATE BOARD OF EDUCATION

P. Paul Ricci,
President
S. David Brandt,
Vice President
Jack Bagan
William Colon
Anne S. Dillman
Ruth H. Mancuso
E. Constance Montgomery
Katherine Neuberger
Sonia B. Ruby
Jack Slater
Susan N. Wilson
Robert J. Wolfenbarger
Timothy Weeks

EX-OFFICIO MEMBER

T. Edward Hollander
Chancellor

SECRETARY

Fred G. Burke

STATE OF NEW JERSEY
DEPARTMENT OF EDUCATION
225 WEST STATE STREET
TRENTON, N J 08625

OFFICE OF THE COMMISSIONER
Post Office Box 2019

Dear Colleague:

Effective educational programs in local school districts are the result of extensive planning and the willingness on the part of administrative and instructional staff to evaluate and modify programs to meet local goals and objectives. A comprehensive testing plan is an integral part of such a program. By focusing on actual needs and requirements, a local district testing plan can facilitate the collection and reporting of information without burdening staff and students with unnecessary or duplicative testing.

As a measurement device, tests can provide a considerable amount of useful information; they can also be misused to an equally enormous extent with particularly disastrous results for those who are affected by inaccurate or misleading interpretations of inappropriate testing.

This document represents the commitment of the New Jersey State Department of Education to guide and assist local school districts in planning, implementing and evaluating a comprehensive testing program which will focus attention and resources in the areas of greatest need.

Sincerely,

Fred G. Burke

Fred G. Burke
Commissioner

State of New Jersey

DEPARTMENT OF EDUCATION
225 WEST STATE STREET
P.O. BOX 2019
TRENTON, NEW JERSEY 08625

DIVISION OF OPERATIONS, RESEARCH AND EVALUATION

Dear Colleague:

A Guide for Local District Testing Programs is the culmination of a two year effort by the Office of State Educational Assessment of the New Jersey State Department of Education. This project was initiated in response to numerous requests from school and district personnel who have become increasingly involved in gathering and reporting test-based information.

As local districts face increasing test-based data requests from numerous sources, the allocation of time, cost and personnel associated with educational testing must be carefully planned.

This Guide is, however, more than a step-by-step planning document for those responsible for coordinating a district-level testing program. It presents, in a single document, the state's testing requirements, a brief introduction to the various aspects of test selection, scoring, analysis and reporting, and a comprehensive bibliography of additional sources of information on tests and testing.

A special word of thanks is extended to David Pearsall, David Engelhardt, and Kris Stanwood who assisted the State Assessment Office staff by reviewing and commenting on the draft versions of this document.

I believe the Guide will prove to be a valuable source book for local testing personnel. I would appreciate receiving from you any comments about its utility.

Sincerely,

Stephen L. Koffler, Director
State Educational Assessment

New Jersey Is An Equal Opportunity Employer

TABLE OF CONTENTS

CHAPTER		PAGE
	Introduction	1
I	New Jersey Department of Education Criteria for Local Testing Programs	3
	General Testing Guidelines	4
	Test Selection Criteria and Development Guidelines	4
	Test Results Guidelines	5
II	Local District Testing Programs in New Jersey	6
	Key Steps to Follow in Implementing a Local District Testing Program	6
III	Implementation of a Local Testing Program	8
	Organizing a Test Planning and Evaluation Committee (Step 1)	8
	Assessing Test-Based Data Needs (Step 2 and Step 3)	9
	Using Paper and Pencil Tests or Other Assessment Techniques (Step 4)	11
	Locating Test Information (Step 5)	14
	Reviewing and Evaluating the Tests (Step 6)	15
	Aptitude	15
	Achievement	15
	Reliability	17
	Validity	19
	Checklist (Step 7)	21
	Summary	23
IV	Locally Developed Tests	24
	Advantages of Locally Developed Tests (Step 8)	24

Disadvantages of Locally Developed Tests.....24
Steps to Follow When Developing A Test (Step 9)....25

V Test Scores.....27
Scoring Procedures (Step 10).....27
Types of Basic Reports.....28
Types of Test Scores.....29
Summary.....33

VI Test Interpretation.....34
Test Administration (Step 11 and Step 12).....34
Interpreting Test Results (Step 13).....35
Using the Test's Technical Manual.....36
Reporting Test Results (Step 13).....37

VII Evaluating the Local Testing Program (Step 14).....39

Appendices

I Sources of Information on Testing - Bibliography.....42

II Annotated Bibliography.....47

III Major U.S. Publishers of Standardized Tests.....52

LIST OF TABLES

TABLE		PAGE
1	Identifying District Programs for which Test-Based Data Needs are Necessary	10
2	Purposes for Testing	12
3	Sample Test Selection Matrix	13

INTRODUCTION

The purpose of this manual is to provide local educational agencies and county office staff with information and specific guidelines on the development, maintenance and evaluation of local district testing programs.

As we enter the 1980's, those who are interested and involved in the education of children continue to seek answers to questions concerning the quality and efficiency of the educational system. Because of this, school officials are becoming more aware of the necessity and advantages of maintaining a comprehensive data collection, retrieval and reporting system.

Testing is one of the most widely used methods of collecting information about students' academic achievement. Every individual who enters and proceeds through a school system is familiar with it. Yet, understanding who to test, when to test, how to test, what test to use, and how to interpret and use test results requires considerable knowledge and experience as well as common sense.

The misuse of tests and test results can have serious short and long term effects on students. But the knowledgeable, well-planned use of tests can help schools provide an optimal educational experience for all students.

This manual contains state approved criteria for local testing programs, a step-by-step presentation of methods for planning, designing and evaluating testing programs, summary information about tests, scoring procedures, types of scores, methods of reporting and interpreting results and a selective reference

section. The document provides information about paper and pencil tests only, it will not discuss issues related to grades, questionnaires, interview procedures or other assessment methods.

CHAPTER ONE

NEW JERSEY STATE DEPARTMENT OF EDUCATION
CRITERIA FOR LOCAL TESTING PROGRAMS

In New Jersey, all pupils must be assessed annually to identify those who do not meet established proficiency standards (N.J.A.C. 6:8-3.8). In grades three, six, nine and eleven, the New Jersey Minimum Basic Skills tests are used to assess each pupil's level of competency in the minimum basic reading and math skills. To assess students in other grades and in other areas of the curriculum, school districts use a variety of commercial and locally developed tests and procedures. All of the testing and assessment activity which occur in a district constitues that district's local testing program. While the components of a local testing program vary from district to district, all testing programs must meet state approved criteria (N.J.A.C. 6:8-3.4).

The state criteria for local testing programs have been organized into three groups: (1) those dealing with testing in general; (2) those related to test selection and development and (3) those related to the release of scores and other test related information. Many of the criteria are legal or administrative mandates from the New Jersey statutes or administrative code. Others are based on requirements related to the assessment and program evaluation components of state and federal basic skills programs.

General Testing Guidelines

1. All students must be assessed on all state goals at least once every five years (18A:7A-9).

2. All students must be assessed annually in the basic skills (18A:7A-10; N.J.A.C. 6:8-3; 6:8-3.8).

3. There must be a uniform testing program in the district. The same test or battery of tests must be used for all students at one particular functional or grade level (i.e., all fifth grade students must be tested with the appropriate levels of the California Achievement Test).

4. To meet department of education requirements for evaluation, assessment and screening, students must be tested at the appropriate time intervals that correspond to the test's norming dates. If testing is solely for local district purposes, the testing can occur at anytime.

5. All tests must validly, reliably and practically measure student competencies in the areas which reflect school and district goals and objectives. In general, one would require a higher reliability for cognitive achievement tests and a lower reliability with other tests, such as occupational interest inventories.

6. No test should be given unless the results will be used.

7. Any unnecessary or duplicative testing should be eliminated.

8. The New Jersey Minimum Basic Skills Tests measure minimum competencies only. They do not always preclude the need for local school districts to administer other tests necessary to assist in decision-making.

9. Teachers, counselors, administrators, parents and students should be involved in the process of determining the need and utility of the various components of the local district's testing program.

Test Selection Criteria and Development Guidelines

1. Different tests have different purposes. The purpose of each test must be clearly defined.

2. All tests must have documented evidence of validity and reliability based upon standard psychometric procedures as outlined in number 5 above.

3. All normative presentations of data must utilize relevant and contemporary norms.

4. All tests should be of an appropriate difficulty level for each student. The Department of Education encourages functional level testing when sufficient data are available for making appropriate interpretations regarding the test results.

5. Group administered intelligence (IQ) and academic aptitude tests may not be used to satisfy basic skills assessment requirements. They may, however, be used for local guidance purposes, etc. outside of the state provisions.

6. Although paper and pencil tests are the most frequently used format testing procedures, the department of education encourages the use of non-traditional or performance measures, instruments or tests, where appropriate in a district's testing program.

Test Results Guidelines

1. In-service training should be given to all staff involved in utilization of test results.

2. Results of all formal achievement testing programs must be made available to students and their parents or legal guardians. Individual student data may not be released except to those persons authorized by the commissioner (N.J.A.C. 6:39-1.2).

3. Results of all assessments must be made available to the public (except those results protected by confidentiality laws, i.e., individual results) (18A:7A-9).

4. Results of formal testing programs must be included in each student's permanent record file.

The remaining chapters in this manual will focus on the development, maintenance and evaluation of local district testing programs. The chapters will provide explanations and rationale for the criteria for local district testing programs and should serve as a general reference guide for district personnel.

CHAPTER TWO

LOCAL DISTRICT TESTING PROGRAMS IN NEW JERSEY

A local district's testing program typically includes various tests and other measurement procedures that provide information about students in order to:

1. assess individual or group educational needs and strengths,
2. develop individual or group instructional plans aimed at improving student competency,
3. select students for placement in specialized instructional programs designed to improve their competencies (e.g., gifted and talented, remedial mathematics, bilingual, etc.),
4. measure individual or group achievement resulting from instruction,
5. evaluate the effectiveness of a particular instructional sequence or educational program,
6. identify learning (curriculum) areas in need of increased instructional emphasis.

An effectively designed and operated testing program provides the structure for the integration of tests into the total educational process. However, such a program requires extensive planning. A list of sequential key steps related to the development, maintenance and evaluation of effective local testing programs is presented below.[1]

Key Steps to Follow in Implementing a Local District Testing Program

1. Assemble a test planning and evaluation committee.
2. Determine the information needs of all programs and grade levels

[1]Further explanations concerning the entire planning process can be found in Resource Notebook of Information for Assessment and Evaluation: Process for T&E, New Jersey Department of Education, 1977.

3. Identify the purposes for testing, the types of decisions required from the test results, and the specific uses for the test results.

4. Consult test information sources to identify a preliminary list of tests for further, more in-depth, review and evaluation. Recognize that different types of tests are available.

5. Decide which testing methods, paper and pencil type testing or other measurement procedures, would best serve the identified purpose for testing.

6. Begin the test review/evaluation process.

   - Study the content of each prospective test to determine its compatability with the program's objectives.

   - Analyze the reliability and validity of the test in light of its purpose and the types of decisions which must be made based on test results.

7. Use the checklist of test selection considerations (on page 22) to determine the most effective test for the district's purposes.

8. If no commercial test is found to satisfy the test-based data needs, consider developing a test locally.

   - Study the advantages and disadvantages of local test development.

   - Follow the local test development steps (page 26).

9. Select or develop an appropriate test.

10. Determine scoring needs and identify types of scores and analyses required, given the purposes for the testing.

11. Plan for the administration and analysis of the tests.

12. Administer and analyze the tests.

13. Report the test results.

14. Evaluate the testing program.

This chapter presented the key steps necessary for the development of a local testing program. Each succeeding chapter in this manual will further define and explain these key steps.

CHAPTER THREE

IMPLEMENTATION OF A LOCAL TESTING PROGRAM

Organizing a Test Planning and Evaluation Committee (Step 1)

The first and perhaps most critical component of a successful local testing program is the development and use of a test planning and evaluation committee.

Local district testing program planning requires the participation of individuals who are familiar with testing and the data needs and requirements for the district's various programs. The formation of a test planning committee is an efficient strategy for ensuring that every program's test-related needs are considered in determining the components of the district's comprehensive test plan.

Representative input from administrators, program directors, grade-level and department staff, guidance counselors, child study team members, testing specialists, students, board of education members and parents will prevent both the potential omission of essential testing material and the scheduling of unnecessary tests. Also, it will ensure that everyone affected by the testing program is represented.

The test planning committee is necessary during all phases of the testing program. It identifies data needs, recommends test instruments or alternative measurement procedures, monitors the administration of certain tests--particularly large scale, district-wide administrations or first use and experimental testing--and

evaluates the content and success of the testing program on a regular basis. It is crucial that such a committee be established in every district.

Assessing Test-Based Data Needs:[2] (Step 2 and Step 3)

A well-designed testing program is characterized by a balanced approach to the amount and type of testing required. A testing program must allow for all the testing that is necessary to meet essential data needs without imposing unnecessary testing on students or test administrators.

While there are many approaches that can be employed in identifying local district testing needs, one of the simplest involves initial examination of testing needs or requirements by grade, program and purpose.

For each grade level or grade cluster, specify every program which is currently, directly or indirectly, associated with testing needs or requirements. The outline in Table 1 provides a list of programs which may assist in identifying programs which may be in operation in a particular district and for which test-based data needs are required either by law, administrative code or department of education directive.

[2]A test-based data need is defined as information about students, usually obtained through testing, which is necessary for the planning, instructional, managerial, and evaluative aspects of a district's educational programs.

TABLE 1

SOME DISTRICT PROGRAMS WHICH

MAY REQUIRE TEST-BASED DATA

I. BASIC SKILLS PROGRAMS

A. Basic Skills Programs (regular/developmental)

B. Basic Skills Programs/Supplemental (list different programs or funding sources)

1. Title I - Basic, Migrant, Non-native English
2. State Compensatory Education
3. Local remedial
4. Other basic skills programs (Right-to-Read, Title IV and State Compensatory Education Research and Development funded projects, Writer's Workshop)

C. Basic Skills - Special Education

II. OTHER INSTRUCTIONAL PROGRAMS

A. Other instructional programs/Regular

B. Other instructional programs/Bilingual

C. Other instructional programs/Gifted and Talented

D. Other instructional programs/Special Education

E. Other instructional programs/Experimental and Innovative

III. NON-INSTRUCTIONAL PROGRAMS

A. Guidance and counseling

B. Teacher inservice

C. Transportation

D. Nutrition program

To determine whether any testing should be planned for students in each grade or cluster in which a particular program operates, consider the different purposes for which testing might be required. Testing provides information which is used to make decisions about the program or the individual students within the program. In test selection, identifying the specific purpose for testing is essential. Table 2 illustrates some of the possible purposes for testing.

A sample selection matrix is presented to summarize the test planning activities to this point. Table 3 shows the design of the matrix which may be used during initial examination of available tests, prior to final selection.

Using Paper and Pencil Tests or Other Assessment Techniques (Step 4)

Once decision-related information needs have been determined, select the method which is most appropriate for obtaining these data. Not all student-related data needs are best served by formal paper and pencil testing. Surveys, questionnaires, formal and informal in-class observations, checklists, pupil-teacher interviews are among the procedures which might provide more valid and meaningful information to the data user especially in non-cognitive areas. The selection of the appropriate measurement procedure depends on the kind of information needed to address identified data needs.

When the decision has been made to use a test to meet one or more of a district's data needs, the actual test selection process begins. THE FIRST STEP IN THIS PROCEDURE IS RECOGNIZING THAT THE

TABLE 2

PURPOSES FOR TESTING

I. TO MAKE DECISIONS WHICH AFFECT INDIVIDUAL STUDENTS

A. Screening

1. Physical examination
2. Psychological examination
3. Psychomotor examination
4. Reading readiness

B. Selection/Placement

1. School entrance
2. Program entrance
3. Class/section entrance

C. Individual Needs Assessment/Diagnosis

1. Cognitive
2. Affective
3. Psychomotor

D. Assessment of Individual Pupil Progress

1. Formative evaluation
2. Adjustment of instructional programs to meet individual needs
3. Grading

E. Individual Pupil Evaluation

1. Summative evaluation
2. Program exit
3. Grade promotion
4. Graduation requirements

II. TO MAKE DECISIONS WHICH AFFECT THE SCHOOL'S PROGRAMS

A. Program Planning

1. Group needs assessment

B. Program Improvment

1. Interim evaluation of group progress
2. Formative evaluation of program implementation

C. Program Evaluation

1. Summative appraisal of program's effectiveness on basis of group's performance

TABLE 3

SAMPLE TEST SELECTION MATRIX*

Grades	Four-Six	Four-Six
Program	School Lunch/Nutrition	Local Remedial Math Program
Purpose	Group Needs Assessment	Diagnosis
Decision	Determine the nutritional needs of students in order to consider the possible establishment of before-school breakfast program	Assignment of students, according to need, to appropriate level of individualized basic math program
Test/Other Measurement Technique	Develop Survey	Paper & Pencil Type Test

*The matrix can be expanded to include the other test selection procedures and test planning considerations discussed in subsequent chapters of this manual.

BEST TEST IS ONE THAT MEETS THE SPECIFIC DATA NEEDS OF THE DISTRICT BY PROVIDING RELIABLE AND VALID INFORMATION WHICH DISTRICT PERSONNEL CAN USE AT A REASONABLE COST.

Locating Test Information (Step 5)

With test purpose in mind, it is then necessary to identify several tests which would be likely candidates for inclusion in the testing program. This step involves conducting a systematic review of tests available.

There are several sources of objective information on available tests, including Tests in Print, Buros' Mental Measurement Yearbooks, Thorndike and Hagen's Measurement and Evaluation in Psychology and Education and Anastasi's Psychological Testing. The reference section at the end of this manual contains bibliographical information on each of these documents.

In addition to these sources, all major test publishers produce catalogues and other sources of information about each of their test products and services. A list of the names and addresses of test publishers has been included in the reference section of this manual.

An important criterion which must be considered when selecting or developing a test is that the instrument be as free from all forms of bias as possible. Both evaluative data based on committee input and empirical evidence based on field testing should be considered to eliminate the possibility of bias.

Once a limited number of tests have been identified, specimen sets should be ordered directly from the test publishers.

A copy of the test, the test administrator's manual, the technical manual and other descriptive information are necessary for a thorough evaluation of each test. Some libraries (Rutgers Graduate School of Education and Educational Testing Service Test Collection Library) have test specimen sets available for public examination.

Reviewing and Evaluating the Tests (Step 6)

Generally, two kinds of tests are used by schools to obtain student performance information: aptitude and achievement tests. These tests are distinguishable more by purpose than by content.

Aptitude

An aptitude test is designed to predict how well a student will perform in the future in a particular program, experience or situation requiring prerequisite skills or knowledge. The Scholastic Aptitude Test (SAT) is an aptitude test which is used in predicting the first year's college performance of high school students.

Achievement

An achievement test samples present behavior to provide an indication of how much a student has learned. The California Achievement Test (CAT) and the New Jersey Minimum Basic Skills Test (MBS) are achievement tests.

For achievement testing there are two basic types of tests: norm-referenced (NRT) and criterion-referenced (CRT) tests. While the content of these test types may be quite similar, they differ in the manner in which the scores are reported and used.

Criterion-referenced tests

Criterion-referenced tests (CRT) are used to determine an individual's performance level with respect to a well-defined and specific set of curriculum objectives rather than to the performance of a reference or norm group. Moreover, a CRT has a very definite standard or criterion for successful performance. Since these tests are designed to provide an estimate of an individual's skills or knowledge in a single area, the scope of these tests is usually narrow. The CRT score can be interpreted in terms of the student's proficiency in the areas being tested. CRT's can also be used to measure an individual's growth over time against certain curricular goals.

Norm-referenced tests

Norm-referenced tests, like criterion referenced tests, contain standards of performance. For norm-referenced tests this standard is the performance of the "norm" group against which the individual student's scores are compared. Norm-referenced tests are used to judge whether a student is growing and changing like students of his/her own age. This is in contrast to criterion-referenced tests which judge an individual's status, not within a group, but with respect to a well-defined behavior domain or area. In the development of the norm-referenced test, items are chosen for their power to discriminate, i.e., their ability to distinguish between the performance of individuals with varying levels of ability.

Test content should correspond as closely as possible to the scope of educational skills or knowledge the district desires

to measure. A careful examination of the test instrument by the test planning and evaluation committee (those people most familiar with both the program's curriculum, objectives and the specific purpose for testing) is the most practical and efficient way to accomplish this selection activity.

A number of factors must be considered when evaluating a test for possible inclusion in a testing program. However, the two most essentional measurement-related factors which supersede all other considerations are reliability and validity.

Reliability

Reliability is a measure of consistency or accuracy of the measurement procedure, technique or instrument. There is always measurement error assoicated with the testing process. The reliability coefficient provides an estimate of how much error exists in an educational or psychological test. The greater the reliability coefficient, the smaller the error.

There are three major types of reliability: (1) internal consistency, (2) test-retest and (3) alternate form. Each of the three types of reliability takes into consideration different sources of error in the measurement process; therefore, there is more than one reliability estimate for any particular test. The types of decisions which will be made based on test results should indicate the type of reliability that is most essential for a particular test.

Internal Consistency

Internal consistency is a form of test reliability which can be estimated from the results of a single administration of a test.

By using a split-half reliability formula, it is possible to obtain a measure of the consistency of the test in terms of content sampling. Kuder-Richardson formulas are used to determine the inter-item reliability of a test, another measure of internal consistency. These formulas provide an estimate of how consistent the test and its items are in measuring the behavior domain or content area being tested.

Test-Retest

Test-retest estimates of reliability provide an estimate of how consistently the test measures performance over a period of time. This type of reliability is usually developed from the results of two administrations of the same test over an approved period of time to the same students. The higher the correlation between the two sets of scores, the more reliable the test.

Alternate Form

There are two kinds of alternate form reliability estimates. One provides an estimate of the consistency of measurement between two equivalent forms of the same test administered at the same time to the same students. The other involves two administrations of different, but equivalent, forms of the same test to the same students at two different times. This latter type of alternate form reliability combines the considerations of both internal consistency and test-retest reliability since it considers both item representativeness and consistency or stability of the measurement over time as sources of measurement error.

Procedures for obtaining reliability information for criterion

referenced tests are different than those described above for norm referenced tests. Because of the relatively recent emergence of CRTs, research and information on the psychometric aspects of such tests is constantly evolving. Information concerning the reliability of CRTs can be found in journal articles such as Hambleton, R. K., Swaminathan, H. and Coulson, D. B., "Criterion-Referenced testing and measurement: A review of technical issues and developments." Review of Educational Research, 1978, 48, 1-47.

The test planning committee should use the information provided in the test's technical manual (which is included in the specimen set, or may be ordered separately from the test publisher or distributor) to evaluate the reliability of a test being considered for use. In their analysis, the following should be considered:

1. Determine in advance the type of reliability which is appropriate for the decisions and uses to be made from the test results. Individual student decisions demand a higher reliability in measurement than program evaluation decisions.

2. If all three types of reliability are important, there may be differences among the reliability coefficients reported for each type, since each accounts for a different source of error.

3. Reliability estimates are reported as coefficients which may range from .0 to 1.00. The higher the coefficient, the more reliable the test. Certain kinds of tests by their very nature have greater accuracy in measuring performance and consequently have higher reliability estimates. It is not uncommon to see reliability coefficients of 0.85 and greater on a standardized achievement test.

4. All other things being equal, select the test with the highest appropriate reliability.

Validity

There are several types of validity a test may exhibit which relate to the different types of tests and the various purposes for which the test is given. The three most basic types of

validity are: content, criterion-related and construct.

Content validity

Content validity is concerned with the extent to which the items in a test are representative of the content area being measured. This type of validity is a particularly important consideration for measures of achievement. Content validity, sometimes referred to as rational or logical validity, can be determined by content experts after a careful review of the test items in relation to a set of instructional objectives or content specifications. For example, a test will have content validity for a particular district if the content of the test matches the objectives or curriculum adopted in the district.

Criterion-related validity

Criterion-related validity is essential whenever one is interested in using a test to predict future outcome, such as success in college. This type of validity is determined statistically, in contrast to the judgmental approach of content validity. Here the concern is how well the test correlates with a measure or criterion of success. Criterion-related validity is particularly important for measures of aptitude. Predictive validity and concurrent validity are two forms of criterion-related validity.

Construct validity

Construct validity is concerned with whether a test actually measures a hypothesized trait or construct. Some examples of constructs are: anxiety, clerical aptitude and reading readiness. The construct validity of a test can be based on evidence obtained

over a period of time, from various research studies which assess the accuracy of the test for measuring the construct in question.

When evaluating a test for validity: (1) remember the reason for administering the test, (2) review the test manuals for the publishers' attempts to assess the test's validity and (3) decide whether a test has sufficient validity for the purpose for which it is being selected.

By using the concept of reliability in test selection or evaluation, we attempt to ascertain how accurate or consistent is the measurement procedure or instrument. In judging the validity of a test, we seek to determine what that instrument is actually measuring. While these two measurement properties are distinct, they are related. If a test is not reliable, it cannot be a valid instrument; however, a reliable test is not necessarily a valid one.

Checklist (Step 7)

While reliability and validity are the essential considerations for selecting a test, several other factors must also be considered. The following checklist may be useful in selecting a commercial test:

Norming Information

_____ Are the norms appropriate for the interpretations you want to make?

_____ Are they recent and up-to-date?

_____ Are they representative of your target population?

_____ Have the norms been developed for testing at various times of the year?

Levels and Forms

_____ Are there levels of the test suitable for all members of the target population?

_____ Does the test include a locator or routing test to help determine the appropriate level of the test for each student?

Content

_____ Do the objectives measured by the test match the objectives of the district?

_____ How much overlap is there between the material on the test and the content covered in the district? (These are actually assessing the content validity of the test at the district level.)

Quality of Materials

_____ How interesting are the test materials?

_____ Are the directions adequate?

_____ Are the test questions of high quality?

_____ Are illustrations, graphs, diagrams within the test printed clearly?

Administration and Timing

_____ Can the test be administered by district personnel?

_____ How much training is involved for administrators?

_____ Can the test be administered to groups of students?

_____ How much time is needed to administer the test, both preparation and actual testing time?

_____ Can the test be administered in sections or must the total test be given during a single session?

Scoring

_____ Can the test be scored by district personnel?

_____ How much training is necessary?

_____ Are tables available to convert raw scores?

_____ Can the test be machine scored at an affordable price?

Costs

_____ What is the cost?

Each test considered should be evaluated separately using the checklist. Checklist results should then be scrutinized to assist in the determination of the most efficient instrument for your particular purposes.

Summary

Test planning procedures continue with a shift from determining whether testing is the appropriate method of data collection to consideration of test characteristics and attributes. Test content must be the primary concern of those involved in the selection of measurement instruments. The scope and level of skills and knowledge covered on the test must closely match the curriculum or educational objectives indicated by the purpose for testing and the decisions which must be made based on the test results.

Identifying the appropriate type of test is the next step considered in this chapter. Aptitude, achievement, norm-referenced and criterion-referenced tests are discussed. The chapter contains a test planning matrix which supplies the framework for the selection process.

Finally, a commercial test selection checklist is presented to provide assistance in selecting a commercial test.

CHAPTER FOUR

LOCALLY DEVELOPED TESTS

It is quite possible that after a detailed review of the commercial testing products available, a district will consider developing its own test to serve its particular data needs. The decision to purchase or develop a test is not an easy one, and there are no definite rules to follow. However, there are certain advantages and disadvantages associated with locally developed tests that should be considered early in the decision-making process.

Advantages of Locally Developed Tests (Step 8)

1. They are custom-made. Local tests can be developed to specifically measure those skills and knowledge which are a part of the local curriculum and which fit the testing purpose.
2. They can be administered at any time of the year. There would be no need to have test dates correspond to commercial publishers' norming dates.
3. They can be as limited or broad as desired in terms of test content and can include as many items per objective as needed to make well-informed decisions.
4. The method of reporting test information can be tailored by the district to meet its special needs.
5. District personnel may feel a stronger sense of acceptance and commitment to a testing program if they have been involved in designing and developing the instrument themselves.
6. The test results would be more appropriately used for instructional purposes because of the test's relevance to the local curriculum.

Disadvantages of Locally Developed Tests

1. Hidden development costs make local tests more costly than generally assumed.

2. Large amounts of development time are required to properly construct and to field test an instrument.

3. A staff person or consultant with testing, measurement and statistical expertise must be available during the test development process.

4. Locally developed tests may not be as generally accepted by the public as commercially available tests.

5. The interpretation of local test results is limited to intra-district comparisons.

6. Multiple forms must be developed for future retesting.

7. Without anchoring the locally developed test to a norm-referenced test, other districts would be unable to interpret the test score of transfer students. Also, without the anchoring, the test could not be used for evaluating Basic Skills Preventive and Remedial Programs.

If, after careful consideration of this information and other locally relevant factors, a district chooses to develop its own test, two new options must be weighed: (1) Should the district develop the test internally using district staff and resources? or (2) Should the district hire an outside contractor to prepare a test according to identified needs and specifications? Regardless of the procedure selected, the following chart outlines the necessary steps for developing a test.

Steps to Follow When Developing a Test (Step 9)

1. Define the attribute and objectives to be measured.

2. Delineate the purpose for the test--What decisions are to be made with the resulting score(s)?

3. Identify the constraints--e.g., budget, resources, testing conditions, etc.

4. Develop a blueprint of the instrument--a list of the content specifications.

5. Determine how many items per specification are necessary.

6. Decide on the format for the test items. What is the response a student will be required to make? What is the stimulus?

7. Decide how to score the items.

8. Write the items. Prepare about twice as many items as would be needed for the final test instrument. Edit items.

9. Have appropriate staff review the items to determine the appropriateness of the items.

10. Have representatives from all relevant minority groups review the items to identify and eliminate possible sources of bias in the items.

11. Develop a procedure for trying out the items on a sample of students comparable to (or part of) your target population.

12. Develop a manual for the test which would include directions for administering the test, scoring the test, interpreting the scores, and other necessary information for test users.

13. Conduct a field test for the items and manuals.

14. Select items for inclusion in the test, based on empirical analyses of the results of the field test.

15. Estimate the reliability of your instrument.

16. Gather evidence for the validity of your instrument for each of the decisions you will make with it.

In summary there is considerably more to developing a valid and reliable instrument than many people believe. Needed are sufficient prior planning, involvement of appropriate people and considerable time. Furthermore, as part of the planning process, state and federal requirements must be considered so that the final product meets those mandates.

However, there are sufficient advantages for local districts with the capability to develop a local testing program. Assistance in the development of these tests can be obtained from the Office of State Educational Assessment or the EICs.

CHAPTER FIVE

TEST SCORES

Test program planning does not end with the selection of the tests to be included in a district's comprehensive testing program. Test results must be converted into test scores which allow for their meaningful, accurate interpretation and use. Successful test planning requires arranging for test scoring. It means understanding the advantages and limitations of the various types of scores available. The selection of a test depends, in part, on the availability of the types of scores and reports the district needs.

Scoring Procedures (Step 10)

One practical aspect of a testing program, often overlooked or taken for granted, is scoring the test. Locally developed tests as well as some commercial tests administered to individual students or small groups of students, are most quickly, efficiently and inexpensively scored by hand.

Answer keys, scoring masks, stencils, self-scoring student answer sheets (for use with specially treated pencil or crayon), or a two-ply student answer sheet with carbon (student responses are automatically transferred to the second page of the answer sheet where the correct answers are pre-recorded) are the most frequently used hand scoring aids. However, when testing large numbers of students, hand scoring becomes an impractical option. Besides being time consuming and more prone to human error, hand

scoring also has the disadvantage of generating only raw-scores which usually require further analysis before the scores can be interpreted or reported.

Automated data processing for test scoring, score conversions and analysis and report generating is essential for most district testing programs. While the majority of such processing is done by a commercial test publisher or supplier, there is a growing trend toward using the local district's computer system for additional test-related processing and storage purposes. In consideration of this, some companies offer punched cards or magnetic tapes in addition to basic reports for districts with automated data processing capabilities. Additionally keypunching, optical scanning and timesharing services can be purchased from a wide variety of companies by districts with computer capabilities.

When the decision is made to use the services of a commercial test publisher, several options became available. As part of a basic test scoring and reporting "package," most publishers provide numerous reports. The following are some types of reports and other test information which are commonly available.

Types of Basic Reports

I. Individual Student Reports

- -test scores (raw and converted)
- -verbal description of performance
- -instructional level
- -individual diagnostic report (objectives mastered and not mastered)
- -individual study guide keyed to test items

II. Class Report

- -alphabetical list of scores (raw and converted)
- -rank order list of scores (raw and converted)

-pupil/objective analysis
-group instructional analysis

III. Building and/or District Summary Reports

-average scores and range by class
-stanines by grades
-percentiles
-normal curve equivalent scores

In addition to the basic score reports many publishers offer a variety of special scoring and reporting services. While the cost for these is usually not included in the basic score report package price, when such services are needed, a commercial service is often most accurate, timely and cost effective. Some optional test services are:

-Item analysis
-Student/parent report form
-Student list/grouping by domain according to mastery or NCE score
-Locator test summary
-Cross reference to texts
-Estimation of probable success
-Local norms
-Frequency distributions in chart or graph form
-Special on-off level frequency distributions with percentile ranks
-Special on-off level group item performance record
-Special percentile ranks/special subpopulation norms
-Pre-post evaluation services
-Administrators summary
-Evaluators summary - summary of means, meadians, quartile distributions, in chart or graph format
-Pupil instructional analysis
-Parents guide to understanding test student/bulletin
-Equivalent scores from one form of a test to another

A comprehensive description of the services that are available can be obtained from the publishers.

Types of Test Scores

There are many types of test scores which are frequently used to report educational test results. Some of the more common are

raw scores, percentiles, stanines, grade-level equivalents and Normal Curve Equivalent (NCE) scores. Percent correct or mastery scores are most commonly associated with criterion-referenced, non-normed test score reporting. The advantages and limitations of each are presented below.

Raw Scores

Raw scores are frequently used to report the results of teacher-made examinations and other similar measures of student achievement. A raw score is equal to the number of correctly answered test items.[1] Aside from the ease with which they may be calculated, raw scores present few advantages for interpretation of test performance. Limited information is provided when a raw score is presented in isolation from other data. For example, a raw score of 53 means nothing by itself. The score could signify drastically different levels of achievement for tests with varying number of items, varying item difficulties, varying comparison group ability, etc. These and other factors must be considered when determining the significance of a raw score.

Percentile Rank

A percentile rank is equal to the percentage of students who score at or below a given score. The use of percentiles facilitates comparisons of student performance in terms of the student's peers.

[1]In some instances where there is a penalty for guessing, the raw score is not simply the number of correctly answered items. Rather, it is the number of correct answers minus the number of wrong answers divided by the number of answer choices minus one.

A student's percentile rank can vary greatly according to the reference group used in calculating that percentile, therefore, it is essential to qualify percentile ranking in terms of specific group affiliation. This is necessary because relative standing varies with the ability of the group used for comparison.

The major limitation of using percentiles is that the distance between points is not equal on all parts of the score scale. Typically percentiles overemphasize differences near the median and under-emphasize differences near the ends of the score scale. Percentiles may be advantageously used to interpret test performance as long as several cautions are heeded. First, a difference of several percentiles should be given greater consideration at the extremes of the distribution than near the middle. Second, percentile ranks must not be added, subtracted, multiplied or divided.

Stanine Scores

Stanine scores were desinged to represent ranges of performance rather than specific values. The word stanine is derived from the phrase--standard score of nine units. Each of the nine stanines, with the exception of stanines one and nine, equals one-half standard deviation in width. The mean average stanine is five (the midpoint of the middle stanine). Because of the nature of stanines, the interpretations that can be made using such scores is not as precise as with other scores.

NCE

The Normal Curve Equivalent (NCE) was developed to provide

a score metric composed of equal-size units. Normal Curve Equivalent units have a mean of 50 and a standard deviation of 21.06. An NCE of 50 matches the 50th percentile of any nationally normed test. NCEs represent an improvement over percentiles in that they are composed of intervals which are equal at all points of the measurement scale. Consequently, a gain of eight NCEs represents the same amount of improvement for students at all points in the score distribution. Another important distinction between NCEs and percentiles is that the equal interval nature of the NCE scale allows for aggregation and averaging of NCEs. Additional information concerning NCEs is available in many of the pamphlets describing the E.S.E.A. Title I evaluation models.

Grade Equivalent Scores

In the past, grade equivalents were among the most commonly used scores when reporting individual performance in terms of normative data. This type of score consists of two digits (e.g., 5.3); the first indicates the grade and the second digit indicates the month. A grade equivalent represents the average score that students at a given grade level achieve. Psychometricians have become increasingly critical of the use of the grade equivalents because they lead to serious misinterpretations of performance. Similar to percentiles, grade equivalents can not be added, subtracted, multiplied or divided.

Percent Correct

Though not commonly used in reporting the results of norm-referenced assessments, the percent correct score is preferred

where individual student performance is considered to be independent of how other students perform. While in years past it has been used most frequently in scoring teacher made tests, the advent of criterion-referenced testing has led to an increase in the use of the percent correct score. The percent correct score is most clearly understood as one's score per 100 items.

Scaled Scores

Scaled scores are developed by transforming the raw score scale to a scale with a desired mean and standard deviation. The transformation can either be a simple linear one or a "normalized" or area transformation. The advantage of "normalized" transformations is the assumption of a resulting scale which is an equal interval scale. An example of a simple linear transformed scale is the Scholastic Aptitude Test scale with a mean of 500 and a standard deviation of 100.

Summary

The scoring of tests is the primary link between test taking and the use of the results. Test program planning must include determination of scoring method--computer or hand scoring using internal resources or commercial scoring services and the types of scores needed to meet utilization and reporting requirements. It may be necessary for a district to obtain results using a variety of scores for varying purposes.

CHAPTER SIX

TEST INTERPRETATION

In order for test results to be meaningful and useful, they must be interpreted. The process of interpreting test results involves the following: understanding and explaining the behaviors, knowledge or skills the test has measured, and (2) making inferences about a student's ability or achievement in a specified area using the test results as one source of information.

Regardless of the test, score mode or method of analysis, the process of interpreting test results should take into consideration certain factors which might limit or increase the validity of the resulting conclusions, including motivation and testwiseness, among others.

Test Administration (Step 11 and Step 12)

A properly selected test which is poorly administered will not provide valid information. Evaluating the administrative aspect of a local testing program involves eliciting information from those directly associated with the testing itself. A test administration comment sheet can be developed and used district-wide for all testing situations. The test administrators could use the comment sheet to report on the following administration variables which may affect test results.

1. Did the student and administrator know that a test was to be administered?
2. Were the administrators and students aware of the purpose of the test?

3. Were test administrators familiar with test material and procedures?

4. Did the examinees understand test directions and response procedures?

5. Were the seating arrangements and the spacing of students adequate?

6. Were the testing sessions too long or too short?

7. Were breaks scheduled at appropriate times?

8. Were interruptions kept at a minimum?

9. Were test make-up periods sufficient?

Test administrators should record comments during or immediately after test administration. The comment sheets should allow for descriptions of any problems which surfaced during the testing sessions as well as suggested adjustments relative to test timing, administration and scoring.

Interpreting Test Results (Step 13)

Presented below are some suggestions and cautions which highlight the factors to be considered in the interpretation process:

1. Keep in mind the purpose and decision for which testing was conducted. Valid interpretations and inferences can only be made if the test from which the results were obtained was designed to provide for such interpretations. For example, a statement about a student's aptitude for advanced mathematics can not be made with reference to a score of 100 on the Minimum Basic Skills Mathematics Test. This test was not designed to provide for that type of decision. On the other hand, it can be validly inferred that a student scoring 29 on the Minimum Basic Skills Reading Test may be in need of remedial assistance in reading.

2. Interpretation of test scores is best made with explicit reference to a mastery level or performance norm. A criterion-referenced test score is defined according to individual mastery standards (correctly answer 75 percent of the items), or group mastery standards (80 percent of all the students must achieve mastery as defined by the individual pupil mastery level). A norm-referenced test score is only meaningful when analyzed in relation to the performance of the appropriate norming population. For example, a score of 97 on a nationally normed test may place the student at the 61st percentile of all students nationwide who took that test. If the test also has local norms, that same score of 97 may place the student at only the 37th percentile of students in the school district who were tested.

3. Inferences and generalizations from the test scores can only be as reliable as the test itself. The reliability of the test instrument is one of the most crucial factors in interpreting test scores. The results of a test which has a relatively low reliability coefficient can not be reliably used as an indicator of the true ability of a particular pupil or group.

4. Special attention must be given to interpreting scores of individuals, particularly when trying to relate subtest scores with composite scores, using combinations of test scores or comparing one subtest score to another (for the same pupil).

5. When possible, test interpretation should take into consideration other valid sources of information about the ability or achievement of the individual or group tested.

6. Exceptional conditions or circumstances affecting the individual or group being tested must be taken into account when interpreting the results, particularly when those conditions produce results which are obviously discrepant when compared to other sources of information about the student or group's performance in the area.

   Factors such as illness, emotional upsets or class interruptions during the test administration must be acknowledged. In some cases, if the disruption is significant, it may be wiser to disregard the test score rather than try to quantify the effect of the disruption on the students' test performance.

7. Interpretations of test scores should take into consideration the standard error of measurement. This is an estimate of the amount of error associated with an individual's score. In most cases, individual scores should be interpreted with a $\pm$ 1 standard error of measurement band around them. This is especially important when looking at differences in scores. A difference between two scores which is less than 1 standard error of measurement is not significant.

8. Information relative to interpretation of the Minimum Basic Skills tests can be found in the two Guidelines booklets published annually.

Using the Test's Technical Manual

An essential tool which should be readily available and frequently consulted during test score interpretation is the technical manual provided by the test publisher. A copy of the test instrument and test administration manual are also helpful. The technical manual, in particular, will assist in answering many of the questions which will guide the interpretation proceedings:

1. What was the test designed to measure?

2. What are the types of norms which can be used for the population tested?

3. What is the estimated reliability of the test (i.e., how much confidence can be placed in the results of the test)?

4. What are the limitations of the uses of the test for making particular decisions about particular populations?

Reporting Test Results (Step 13)

An essential part of reporting test results is considering the audience or users of the results and their particular needs. The reporting of test data requires a variety of skills including the ability to communicate effectively in written, graphic and/or oral form; being able to anticipate the user's level of testing sophistication; and knowledge of the inferences that can and can not be made validly about pupil or group performance based on the test's own limitations.

Different strategies can be employed to produce meaningful test score reports for different audiences. Test results which are necessary for teachers to make instructional decisions would certainly differ from test score reports for the local school board. The difference between these two would be in terms of:

1. level of specificity--individual student score vs. aggregated scores

2. level of aggregation--classroom or district-wide scores

3. amount and focus of accompanying materials such as narrative interpretations and explanation of terms used

4. type of score reported--raw score and converted scores vs. average scores, stanines, ranges, etc.

Particular attention should be paid to the timing of the release of public information about test results. Students and

parents should receive individual test information before district-wide results are made public. It is also necessary to prepare and have available information which will assist those unfamiliar with a particular test or score mode in making accurate and appropriate inferences from the test results.

CHAPTER SEVEN

EVALUATING THE LOCAL TESTING PROGRAM (Step 14)

A testing program should provide much of the information needed to meet a district's curriculum, evaluation and reporting requirements. To ensure that the testing program satisfies the district's data needs, the test planning and evaluation committee or a representative cross section of those involved in the district's instructional and administrative functions should take part in the evaluation of the testing program.

The first evaluation requirement is really part of the test selection procedures described in chapter three. Each test must be reviewed to determine if it is adequate and addresses the purpose for which it was administered. After a careful review of all tests administered, any test whose results were not used should be considered unnecessary and should be removed from the testing program.

In addition to reviewing the test to decide if it was necessary, each test must be examined to determine if its selection was appropriate. When reviewing a test to judge if it is the best instrument for collecting specific data, the following factors should be considered:

1. validity
2. reliability
3. appropriate norms
4. suitable length
5. appropriate vocabulary and reading difficulty levels
6. timelines and adequate scoring methods

7. adequacy of the receipt of test results to meet local, state and federal reporting requirements
8. viable administration procedures
9. suitable cost

At the end of the year, or when all testing is completed, the evaluation of the entire testing program should be conducted. Test results and comment sheets from all test administrators should be available to the test program evaluation committee or other representatives of those groups involved in the testing process. It is helpful, but not necessary, to have a psychometrician present who can give in-depth statistical interpretation, norm information and measurement advise (some districts may have personnel with sufficient knowledge to perform those activities).

The best method for assuring that the test-based data needs of a district are being met and that no unnecessary, unused or redundant testing is present, is to plan the district-wide testing program in advance. However, a thorough testing program evaluation is still necessary to pinpoint problems such as duplication of testing, too much testing, inappropriate test selections, administration problems, poor use of test results and lack of continuity of the program.

Some questions which might guide a comprehensive review of the testing program are:

1. Is the test planning committee functioning smoothly? Does it have sufficient information and cooperation to perform its duties?
2. Are tests being selected according to good test selection criteria (e.g., is a Testing Planning Matrix or similar procedure used)?
3. Are all district testing needs and requirements met?
4. Is there any unnecessary testing?

5. Do test administrations run smoothly?

6. Are all test results scored, analyzed, interpreted and reported accurately and promptly?

7. Are all test results used for the purpose for which they were given?

8. Were results recorded or retained for future use?

9. Are there any other (new) testing requirements which must be considered?

10. Are teachers, parents, administrators and board members aware of the testing program and how it functions?

11. Are there new, improved testing procedures or instruments which should be considered for the local testing program?

In setting up test program evaluation sessions, the following suggestions may be helpful:

1. Include representatives from all groups involved in the testing program in the evaluation procedure.

2. Have test materials available for reference.

3. Keep printouts of the results for review.

4. Collate input from all examiners regarding test administration procedures.

5. Compile a list of facts and figures about validity and reliability of the tests.

6. Maintain, for review, a list of how test results were utilized.

7. Provide readability and difficulty levels of tests for review.

8. Have available a tabulation of costs of each test and of the program in its entirety.

9. Have available, for examination, the time line of processes involved from time of test administration to use of test results.

APPENDIX I

Sources of Information on Testing

BIBLIOGRAPHY

The following bibliography includes test related books, journals, and reports.

Ahman, J.S., & Glock, M.D. Evaluating Pupil Growth, Principles of Tests and Measurements. (4th ed.) Boston: Allyn and Bacon, Inc., 1971.

Aiken, L.R. Jr. Psychological and Educational Testing. (3rd ed.) Boston: Allyn and Bacon, Inc., 1978.

Anderson, S.B., Ball, S., Murphy, R.T., et al. Encyclopedia of Educational Evaluation. San Francisco: Jossey-Bass, 1975.

Angoff, William H. " Scales, Norms and Equivalent Scores," Educational Measurement Washington D.C. : American Council on Education, 1971.

Ashman, J. Stanley. Testing Student Achievements and Aptitudes. Washington Center for Applied Research in Education, 1962.

Bloom, B.S. Taxonomy of Educational Objectives: Cognitive Domain. New York: David McKay, 1956.

Borg, Walter R. and Gall, Meredith D., Educational Research - An Introduction New York: David McKay Company, Inc., 1976.

Borich, G.D. (Ed.) Evaluating Educational Programs and Products. Englewood Cliffs, N.J. : Educational Technology Publications, 1974.

Brown, F.G. Measurement and Evaluation. (1st ed.) Itasca, Ill.: F.E. Peacock Publishers, Inc., 1971.

Brown, F.G. Principles of Educational and Psychological Testing. Hinsdale, Ill.: The Dryden Press, Inc., 1970.

Bruno, Nancy; Campbell, Don; and Schabacker, William; Statewide Assessment: Methods and Concerns, Center for Statewide Educational Assessment, Educational Testing Service, Princeton, New Jersey.

Chauncey, Henry and John E. Dobbin. Testing : its Place in Education Today(1st ed.) New York:Harper & Row, 1963.

Cronbach, L.J. Essentials of Psychological Testing. (2nd ed.) New York: Harper & Row, 1960.

Cureton, E.E., and Tuley, J.W. " Smoothing Frequency Distributions, Equating Tests, and Preparing Norms," American Psychologist, vol.4 (1951), 404.

Engelhart, M.D. Improving Classroom Testing, Washington: National Education Association, 1964.

Ebel, R.L. Measuring Educational Achievement. Englewood Cliffs, N.J. : Prentice Hall, 1965.

Feldt, L.S. "What Size Samples for Methods/Materials Experiments? " Journal of Educational Measurement, 1973, 10, No. 3, 225.

Flanders, N.A. Teacher Influence, Pupil Attitudes and Achievement. Cooperative Research Monograph No. 12 OE 25040. Washington, D.C.: U.S. Government Printing Office, 1965.

French, J.E., and W.B. Michael. Standards for Educational Psychological Tests and Manuals. Washington: American Psychological Association, 1966. Revised edition - Standards for Educational and Psychological Tests. Washington: American Psychological Association, 1974.

Good, T.L., & Brophy, J.E. Looking in Classrooms. San Francisco: Harper & Row, 1973 Reprinted c 1975, Harper & Row, Publishers.

Gronlund, N.E. Preparing Criterion-Referenced Tests for Classroom Instruction, A Title in the Current-Topics in Classroom Instruction Series. New York: The MacMillan Company, 1973.

Gronlund, N.E. Measurement and Evaluation in Teaching. New York: The MacMillan Company, 1971.

Gronlund, N.E. Constructing Achievement Tests. New Jersey: Prentice Hall, Inc., 1968.

House, E.R. (Ed.) School Evaluation: Politics and Process. Berkeley: McCutchan, 1973.

Isaac, S., & Michael, W.B. Handbook in Research and Evaluation. San Diego: Robert R. Knapp, 1971.

Karmel, Louis J. Testing in Our Schools. New York: MacMillan Publishing Company, 1966.

Krathwohl, D.R., Bloom, B.S., & Masia, B.B. Taxonomy of Educational Objectives: Affective Domain. New York: David McKay, 1964.

Krejcie, R.V., & Morgan, D.W. "Determining Sample Size for Research Activities." Educational and Psychological Measurement, 1970, 30, 607-610.

Kropp, R.P., & Verner, C. "An attitude scale technique for evaluating meetings." Adult Education, Summer 1957, VII (4).

Lehman, Irvin J. and Mehrens, William A. Measurement and Evaluation in Education and Psychology. New York: Holt, Rinehart and Winston, Inc., 1973.

Lennon, R.T. " Equating Nonparallel Tests," Journal of Educational Measurement, Vol. 1 (1964), 15 - 18.

Linton, M., & Gallo, P.S. Jr. The Practical Statistician: Simplified Handbook of Statistics. Monterey, California: Brooks Cole, 1975.

Lord, F.M. " A Survey of Observed Test-Score Distributions with Respect to Skewness and Kurtosis, " Educational Psychological Measurement, Vol. 15 (1955) 383-389.

Loret, P.G. et al. Anchor Test Study: Equivalence and norms tables selected for reading achievement tests (grades 4,5,&6. Office of Education Report 74-305, U.S. Government Printing Office, Washington, D.C., 1974.

Lyman, Howard B. Test Scores and What They Mean. New Jersey: Prentice Hall, Inc., 1971.

Marshall, J.C., & Hales, L.W. Essentials of Testing. Menlo Park, California: Addison-Wesley Publishing Company, 1972.

Marshall, J.C., & Hales, L.W. Classroom Test Construction. Menlo Park, California: Addison-Wesley Publishing Company, 1971.

McFarland, S.J., & Hereford, C.F. Statistics and Measurement in the Classroom, Psychological Foundations of Education Series. (2nd ed.) Iowa: William C. Bloom Company, Publishers, 1971.

McLaughlin, K.F. Interpretation of Test Results. Washington: U.S. Government Printing Office, 1964.

Mehrens, W.A., & Lehmann, I.J. Measurement and Evaluation and Psychology. New York: Holt, Rinehart, and Winston, Inc., 1973.

Meissner, Frederick Minimum Basic Skills Program-Minimum Basic Skills Advisory Committee's Final Report, January 12, 1978.

Metfessel, N.S., & Michael, W.B. A Paradigm Involving Multiple Criterion Measures for the Evaluation of Effectiveness of School Programs. Educational and psychological Measurement, 1967.

Nedelsky, L. " Absolute Grading Standards for Objective Tests," Educational and Psychological Measurement, 1954, 14, No.1.

Nunnally, J.C.Jr. Introduction to Psychological Measurement. New York: McGraw-Hill Book Company, 1970.

Nunnally, J.C. Educational Measurement and Evaluation. New York: McGraw-Hill Book Company, 1964.

Payne, D.A. (Ed.) Curriculum Evaluation. Lexington, Mass.: D.C. Heath, 1974.

Popham, W.J. Educational Evaluation. Englewood Cliffs, N.J. : Prentice Hall, 1975.

Provus, M.M. Discrepancy Evaluation: For Educational Program Improvement and Assessment. Berkeley: McCutchan, 1971.

Psychological Corporation. Test Service Bulletin # 48. Printed by permission of the publisher, Psychological Corpporation.

Simpson, E.J. The Classification of Educational Objectives, Psychomotor Domain. Illinois Teacher, 1966-67.

Simpson, E.J. Scheme for Classification of Educational Objectives: Psychomotor Domain. Unpublished manuscript, 1969.

Slonin, M.J. Sampling: A Quick, Reliable Guide to Practical Statistics. New York: Simon & Schuster, 1967, 74.

Small-sample Techniques. The NEA Research Bulletin, December 1960, 38, 99-104.

Smith, F.M., & Adams, S. Educational Measurement for the Classroom Teacher. (2nd ed.) New York: Harper & Row, Publishers, 1972.

Stake, R.E. The Countenance of Educational Evaluation. Teachers College Record, 1967, 68, 523-540.

Swineford, F., and Fan, C. " A Method of Score Conversion Through Item Statistics," Psychometrika, vol. 22 (1957) 185-188.

Swineford, F. Statistical Report, New Jersey Educational Assessment Program, Special Score Equating Study. Princeton: Educational Testing Service, 1977.

Thorndike, R.L., & Hagen, E. Measurement and Evaluation in Psychology and Education. (3rd ed.) New York: John Wiley and Sons, Inc., 1969.

Thurstone, L.L. " A Method of Scaling Psychological and Educational Tests," Journal of Educational Psychology, vol. 16 (1925) 433-451.

Tyler, L.E. Tests and Measurements. (2nd ed.) Englewood Cliffs, N.J.: Prentice Hall, Inc., 1971.

Tyler, R.W., Gagne, R.M., & Scriven, M. Perspectives on Curriculum Evaluation. American Education Research Association Monograph #1 in Curriculum Evaluation Series. Chicago: Rand McNally, 1967.

Tyler, R.W. (Ed.) Educational Evaluation: New Roles, New Means. Chicago: University of Chicago Press, 1969.

Tyler, R.W., & Wolf, R.M. Crucial Issues in Testing. Berkeley: McCutchan, 1974.

Walberg, H.J. Evaluating Educational Performance. Berkeley: McCutchan, 1974.

Walker, H.M., & Levy, J. Statistical Inference. New York: Holt, Rinehart and Winston, 1953.

Webb, Louise W. & Others Testing in the Elementary School. Farrar and Rinehart Inc., Publisher, N.Y.C. 1939

Wick, J.W. Educational Measurement, Where Are We Going and How Will We Know When We Get There? Columbus, Ohio: Charles E. Merrill Publishing Company, 1973.

Wilks, S.S. (ed.) Scaling and Equating College Board Tests. Princeton: Educational Testing Service, 1961.

Womer, F.B. Guidance Monograph Series III: Testing, Basic Concepts in Testing. Boston: Houghton Mifflin Company, 1968.

Worthen, B.R., & Sanders, J.R. Educational Evaluation: Theory and Practice. Worthington, Ohio: C.A. Jones, 1973.

APPENDIX 2

ANNOTATED BIBLIOGRAPHY

The following annotated bibliographies are from the Eric Clearinghouse on Tests, Measurement, and Evaluation. Different types of tests include: measures of social skills, measures related to school based attitudes, self concept, educationally disadvantaged, assessment of teachers, and criterion referenced tests.

ED 056 082. Rosen, Pamela, and Horne, Eleanor V. Language Development Tests: An Annotated Bibliography, 1971.

Information on language development tests for pre-school and elementary grades 1 - 3 is presented in this annotation. It includes: purpose of the test, tested skills, competencies, administration, scoring, interpretation, and standardization.

ED 056 083. Guthrie, P.D. and Horne, Eleanor V. School Readiness Measures: An Annotated Bibliography. 1971.

This is a listing of general school readiness measures and gives the age groups for which each is intended. Also, provides information on purpose, test skills, behaviors, administration, scoring, interpretation, and standardization.

ED 056 085. Guthrie, P.D. and Others Measures of Social Skills: An annotated Bibliography. 1971.

Pre-school through 3rd grade level instruments for measuring various social skills. Designed for measuring social competency, social maturity, social sensitivity and attitudes towards others.

ED 074 071. Knapp, Joan, Comp. An Omnibus of Measures Related to School-Based Attitudes. 1972.

These are summaries 16 different instruments used for the purpose of measuring school-based attitudes. Self-report inventories for children between the 4-8 age group and others designed for grades 12-14.

ED 080 534. Knapp, Joan, Comp. A Selection of Self Concept Measures. 1973.

The instruments described here are for the purpose of measuring self-concept. All of the instruments described are suitable for a full range of school age children.

ED 083 318. Rosen, Pamela, ed. Tests for Educationally Disadvantaged Adults. 1973.

This annotated bibliography describes 65 different instruments for measuring achievement and aptitude. Intended for adults with limited education due to learning disabilities and/or educational handicaps.

ED 083 319. Rosen, Pamela, ed. Self-Concept Measures: Grade 7 and Above. 1973.

The instruments described in this test bibliography are measures of self-concept and self-esteem for use in grades 7 and above. The purpose of these tests is to define the individuals perceptions and evaluations of himself in the school environment.

ED 083 321. Rosen, Pamela, ed. Attitudes Toward School and School Adjustment Grades 4-6. 1973.

This bibliography of measures of attitudes toward school and chool adjustment is appropriate for grades 4-6 (also grades 7-12). It includes pupils attitudes toward themselves as learners, the learning process, school environment, school subjects and teachers.

ED 083 322. Rosen, Pamela, ed. Assessment of Teachers. 1973.

This is a bibliography of a variety of instruments used for assessing teachers. Includes are devices to measure knowledge in general and specific areas of education; a) attitude or self-report measures b) student evaluations c) devices used for observing teacher competency and pupil interaction.

ED 139 118. New Jersey Minimum Basic Skills Program. April 77

This paper was presented at the Annual Meeting of the National Council on Measurement in Education. State tests are administered each year to grades 3, 6, 9, and 11. The M.B.S. program is implemented by the N.J. Department of Education for the purpose of assessing students' mastery of minimum basic skills. Guidelines are available to assist local districts in providing basic skills programs.

ED 147 326. District Utilization of the Statewide Testing Program. Occasional Papers in Education. New Jersey Dept. of Education, Division of Research, Planning, and Evaluation. June 77.

The N.J. Educational Assessment Program (EAP) was a statewide testing program which assessed reading and math. This information is disseminated at state and local levels. Use of the data results vary from district to district but include: program changes, modification of reading and math instruction, comparisons of year to year results, administrative changes and inservice workshops.

ED 148 832. Staff Development Through Inservice Workshops in Testing for Teachers, Counselors, and Principals. Nov. 76.

Testing workshps were conducted for teachers, counselors, and administrators in a Fremont, California school. Discussion focuses on test interpretation, administration, analysis and relevancy of test data. Handbooks which include reference materials on testing are included in the appendices.

VII. An Annotated Bibliography of Guides for Test Selection

Compiled by John Jegi, Director, ACCESS Information Center, Contra Costa County Superintendent of Schools Office

Buros, Oscar K., ed. Mental Measurements Yearbooks: 1st ed. - 1938 (re-issued 1972); 2nd ed. - 1941 (re-issued 1972); 3rd ed. - 1949; 4th ed. - 1953; 5th ed. - 1959; 6th ed. - 1965; 7th ed. - 1972 (2 vols.). Highland Park, New Jersey: The Gryphon Press.

Single best source of critical reviews of tests. Each yearbook contains critical reviews of all obtainable published tests and books on measurement written in English. Most publications are reviewed independently by two or more specialists. Reviews in earlier editions are cross-referenced in later ones.

_____. Tests in Print. Highland Park, New Jersey: The Gryphon Press, 1961.

A comprehensive test bibliography and index to the first five books in the Mental Measurements Yearbook series. Each test mentioned includes information concerning test title, appropriate grade levels, publication data, special short comments about the test, number and types of scores provided, authors, publishers, and reference to test reviews in Mental Measurements Yearbooks.

_____. Tests in Print. Volume 2. Highland Park, New Jersey: The Gryphon Press, 1974.

Index to tests still in print that are listed in all seven Mental Measurements Yearbooks. Includes bibliographies of references on the construction, use and validity of specific tests published through 1971.

_____. Personality Tests and Reviews. Highland Park, New Jersey: The Gryphon Press, 1970.

Provides compilation of personality test reviews and specific test bibliographies listed in the first six Mental Measurements Yearbooks. Includes new material on personality testing, including a comprehensive bibliography of 513 personality tests and 7,116 new references dealing with the construction, use, and validity of specific tests. Also includes a master index to the nonpersonality tests, reviews and references to the first six Mental Measurements Yearbooks. Eighty tests--new, revised, or supplemented since the Sixth Yearbook and not listed in the Seventh Yearbook--are included.

_____. Test Collection Bulletin. A Quarterly Digest of Information on Tests. Vol.4 #3 Educational Testing Service, Princeton, NJ July 1970. (Ed 150 193).

This is one of a series of test collection bulletins provided by Educational Testing Service (E.T.S.) which provides information concerning tests acquired by E.T.S., test announcements, dates of testing programs, changes in test distribution, tests no longer available, test reviews, new references (annotated), new publisher listings, test scoring services and systems, and addresses of organizations and publishers they mention in the bulletin.

_____. Reading Tests and Reviews. Highland Park, New Jersey: The Gryphon Press, 1968.

Includes a comprehensive bibliography of reading tests as of early 1968, a reprinting of all reading test reviews in the first six Mental Measurements Yearbooks, and a master classified. index to all other tests and reviews in the first six Yearbooks. Includes information about 33 reading tests--new, revised, or supplemented since the Sixth Yearbook which are not listed in the Seventh Yearbook.

Hoepfner, Ralph, Ed. CSE Elementary School Test Evaluations. Los Angeles: Center for the Study of Evaluation, UCLA, 1970.

This book contains a compendium of tests, keyed to educational objectives of elementary school education, and evaluated by measurement experts and educators for such characteristics as meaningfulness, examineee appropriateness, administrative usability. and quality of standardization.

Hoepfner, Ralph; Stern, Carolyn; and Nummendal, Susan G., eds. CSE-ECRC Preschool/Kindergarten Test Evaluations. Los Angeles: Center for the Study of Evaluation and the Early Childhood Research Center, UCLA, 1971.

This book contains a compendium of tests, keyed to educational objectives of early childhood education, and evaluated by measurement experts and educators.

Johnson, Orval G., and Bommarito, James. Tests and Measurements in Child Development: A Handbook. San Francisco: Jossey-Bass, Inc., Publishers, 1971.

A guide to more than 300 measures of child behavior and development not available from test publishers. Authors cite six criteria for inclusion: (1) suitability for use with children between birth and age twelve; (2) availability to professionals; (3) unpublished, not commercially available; (4) permit development of norms and reliability and validity data; (5) include enough information for effective use; (6) technically useable measures classified in ten categories: (a) cognitive, (b) personality and emotional characteristics, (c) children's perceptions of environment, (d) self-concept, (e) actual environment, (f) motor skills, brain injury, sensory perception, (g) physical attributes, (h) attitudes and interests not otherwise classified, (i) social behavior, and (j) measures not fitting the above categories.

Johnson, T.J., and Hess, R.J. Tests in the Arts. St. Louis: Central Midwestern Regional Educational Laboratory (CEMREL), 3120 - 59th Street, St. Louis, Mo., 1970.

Indexes and abstracts all known measuring instruments and tests applicable to the arts and provides a brief but comprehensive overview of the various psychometric methodologies utilized in the development of the instruments.

Robinson, John P., and Shaver, Philip R. Measures of Social Psychological Attitudes. Ann Arbor, Michigan: Publications Division, Institute for Social Research, University of Michigan, 1969*.

Review of 106 test instruments grouped into eight general categories: Life satisfaction; Self-esteem; Alienation; Authoritarianism; Socio-political attitudes; Values; General attitudes toward people; and Religious attitudes. Evaluation of instruments' psychometric properties given as well as ease of administration and scoring.

*Available in 1974 revised edition.

Simon, A. and E.G. Boyer. Mirrors for Behavior: An Anthology of Classroom Observations Instruments. Philadelphia: Research for Better Schools, 1967.

An annotated compilation of 86 observations instruments representating a variety of approaches, both in the affective and cognitive domains. Extensive bibliography.

Walker, Deborah K. Socioemotional Measures for Preschool and Kindergarten Children. San Francisco: Jossey-Bass, Inc. Publishers, 1973.

Description of 143 tests and measures of social and emotional development including titles and dates of publication or copyright; author; appropriate age range; measurement technique; source in which measure is described; description of the instrument; norms available; validity studies; and reliability evidence.

Wall, Janet, and Summerlin, Lee. Standardized Science Tests: A Descriptive Listing. Washington, D.C.: National Science Teachers' Association, 1973. (Order direct from NSTA, 1201 Sixteenth Street, NW, Washington, D.C. 20036 $1.50).

A compilation of virtually all the standardized science tests published since 1959 available to elementary and secondary science teachers. (57 pages).

APPENDIX 3

MAJOR U.S. PUBLISHERS OF STANDARDIZED TESTS

Addison-Wesley Publishing Company
Reading, Massachusetts 01867
617-944-3700

The American College Testing Program (ACT)
P.O. Box 168
Iowa City, Iowa 52240
319-356-3711

American Guidance Service, Inc.
Publishers' Building
Circle Pines, Minnesota 55014
612-786-4343

The Bobbs-Merrill Company, Inc.
4300 West 62nd Street
Indianapolis, Indiana 46268
317-291-3100

Bureau of Educational Measurements
Kansas State Teachers College
Emporia, Kansas 66801
316-343-1200

Bureau of Educational Research and Service
C-20 East Hall
The University of Iowa
Iowa City, Iowa 52240
319-353-3823

C.H. Stoelting Company
See Stoelting Company

CTB/McGraw-Hill
Western United States:
(Main Office)
Del Monte Research Park
Monterey, California 93940
408-649-8400
Eastern United States:
Manchester Road
Manchester, Missouri 63011
314-227-1600

Committee on Diagnostic Reading Tests, Inc.
Mountain Home, North Carolina 28758
704-693-5223

Consulting Psychologists Press, Inc.
577 College Avenue
Palo Alto, California 94306
415-326-4448

Cooperative Tests & Services
See Addison-Wesley Pub. Co.

Educational & Industrial Testing Service
P.O. Box 7234
San Diego, California 92107
714-488-1666

Educational Records Bureau
Box 619
Princeton, New Jersey 08541
609-921-9000

Educational Test Bureau
See American Guidance Service, Inc.

Educational Testing Service
Princeton, New Jersey 08541
609-921-9000
Western Office:
1947 Center Street
Berkeley, California 94704
415-849-0950
Midwestern Office:
960 Grove Street
Evanston, Illinois 60201
312-869-7700

Follett Publishing Company...A Division of Follett Corporation
Department DM
1010 West Washington Boulevard
Chicago, Illinois 60607
312-666-5855

Ginn and Company
P.O. Box 2649
1250 Fairwood Avenue
Columbus, Ohio 43216
614-253-8661

Grune and Stratton, Inc.
111 Fifth Avenue
New York, New York 10003
212-741-6800

Guidance Testing Associates of
St. Mary's University
1 Camino Santa Maria
San Antonio, Texas 78284
512-436-3304

Harcourt Brace Jovanovich, Inc.
See The Psychological Corp.

Houghton Mifflin Company
Pennington-Hopewell Road
Hopewell, New Jersey 08525
609-466-1950

Institute for Personality and
Ability Testing (IPAT)
1602 Coronado Drive
Champaign, Illinois 61822
217-352-4739

Martin M. Bruce, Publishers
340 Oxford Road
New Rochelle, New York 10804
914-235-4450

McGraw-Hill Book Company
See CTB/McGraw-Hill

Personnel Press
See Ginn and Company

Priority Innovations, Inc.
P.O. Box 792
Skokie, Illinois 60076
312-729-1434

The Psychological Corporation
757 Third Avenue
New York, New York 10017
212-888-3500

Psychological Research Services
Case Western Reserve University
1695 Magnolia Drive
Cleveland, Ohio 44106
216-368-3536

Psychological Test Specialists
Box 1441
Missoula, Montana 59801

Psychologists and Educators, Inc.
Suite 212
211 West State Street
Jacksonville, Illinois 62650
217-243-2135

Psychometric Affiliates
Box 3167
Munster, Indiana 46321
219-836-1661

Richardson, Bellows, Henry and
Company, Inc.
1140 Connecticut Avenue, N.W.
Washington, D.C. 20036
202-659-3755

Scholastic Testing Service, Inc.
480 Meyer Road
Bensenville, Illinois 60106
312-766-7150

Science Research Associates, Inc.
155 North Wacker Drive
Chicago, Illinois 60606
312-984-2000

Sheridan Psychological Services,
Inc.
P.O. Box 6101
Orange, California 92667
714-639-2595

Stoelting Company
1350 South Kostner
Chicago, Illinois 60623
312-522-4500

Teachers College Press
Teachers College
Columbia University
New York, New York 10027
212-678-3991

University Bookstore
Purdue University
360 State Street
West Lafayette, Indiana 47906
317-743-9618

Western Psychological Services
12031 Wilshire Boulevard
Los Angeles, California 90025
213-478-2061